Abstract

Provided is a similarity search method that makes use of a localized distance metric. The data includes a collection of items, wherein each item is associated with a set of properties. The distance between two items is defined in terms of the number of items in the collection that are associated with the set of properties common to the two items. A query is generally composed of a set of properties. The distance between a query and an item is defined in terms of the number of items in the collection that are associated with the set of properties common to the query and the item. The properties can be of various types, such as binary, partially ordered, or numeric. The distance metric may be applied explicitly or implicitly for similarity search. One embodiment of this invention uses random walks such that the similarity search can be performed exactly or approximately, trading-off between accuracy and performance. The distance metric of the present invention can also be the basis for matching and clustering applications. In these contexts, the distance metric of the present invention may be used to build a graph, to which matching or clustering algorithms can be applied.